The Unreasonable Effectiveness of Deep Evidential Regression

Nis Meinert[1]*, **Jakob Gawlikowski**[2]*, **Alexander Lavin**[1]

[1]Pasteur Labs, 19 Morris Avenue, Brooklyn Navy Yard Building 128, Brooklyn, NY 11205, USA
[2]German Aerospace Center, Institute of Data Science, Mälzerstraße 3-5, 07745 Jena, Germany
nis.meinert@simulation.science, jakob.gawlikowski@dlr.de, lavin@simulation.science

Abstract

There is a significant need for principled uncertainty reasoning in machine learning systems as they are increasingly deployed in safety-critical domains. A new approach with uncertainty-aware regression-based neural networks (NNs), based on learning evidential distributions for aleatoric and epistemic uncertainties, shows promise over traditional deterministic methods and typical Bayesian NNs, notably with the capabilities to disentangle aleatoric and epistemic uncertainties. Despite some empirical success of Deep Evidential Regression (DER), there are important gaps in the mathematical foundation that raise the question of why the proposed technique seemingly works. We detail the theoretical shortcomings and analyze the performance on synthetic and real-world data sets, showing that Deep Evidential Regression is a heuristic rather than an exact uncertainty quantification. We go on to discuss corrections and redefinitions of how aleatoric and epistemic uncertainties should be extracted from NNs.

1 Introduction

Using neural networks (NNs) for regression tasks is one of the main applications of modern machine learning. Given a dataset of $(\vec{x}_i, \vec{y}_i)$ pairs, the typical objective is to train a NN $\vec{f}_i \equiv \vec{f}(\vec{x}_i|\boldsymbol{\omega})$ w.r.t. $\boldsymbol{\omega}$ such that a given loss $\mathcal{L}(\vec{y}_i, \vec{f}_i) \equiv \mathcal{L}_i(\boldsymbol{\omega})$ becomes minimal for each $(\vec{x}_i, \vec{y}_i)$ pair. Traditional regression-based NNs are designed to output the regression target, a.k.a., the prediction for $\vec{y}_i$, directly which allows a subsequent minimization, for example of the sum of squares:

$$\min_{\boldsymbol{\omega}} \sum_i \mathcal{L}_i(\boldsymbol{\omega}) = \min_{\boldsymbol{\omega}} \sum_i \left(\vec{y}_i - \vec{f}(\vec{x}_i|\boldsymbol{\omega})\right)^2. \quad (1)$$

Technically, this is nothing but a fit of a model $\vec{f}$, parameterized with $\boldsymbol{\omega}$, w.r.t. $\sum_i \mathcal{L}_i$ to data. As with any fit, the model has to find a balance between being too specific (over-fitting) and being too general (under-fitting). In machine learning this balance is typically evaluated by analyzing the trained model on a separated part of the given data which was not seen during training. In practice, no model will be able to describe this evaluation sample perfectly and deviations can be categorized into two groups: *aleatoric* and *epistemic* uncertainties (Hora 1996; Kiureghian and Ditlevsen 2009; Kendall and Gal 2017; Hüllermeier and Waegeman 2021). The former quantifies system stochasticity such as observation and process noise, and the latter is model-based or subjective uncertainty due to limited data.

The field has largely focused on Bayesian NN approaches that use Monte Carlo sampling and other approximate inference techniques to estimate uncertainty of deep NN models. A different approach to uncertainty-aware NNs may be useful to more efficiency quantify, and also to disentangle, the several types of uncertainties: *Deep Evidential Regression* (DER) aims to simultaneously predict both uncertainty types in a single forward pass without sampling or utilization of out-of-distribution data, based on learning evidential distributions for aleatoric and epistemic uncertainties (Amini et al. 2020). Yet only with simple empirical demonstrations on univariate regression tasks, this technique has already been applied and recommended in medical and other safety critical applications (Liu et al. 2021; Soleimany et al. 2021; Cai et al. 2021; Chen, Bromuri, and van Eekelen 2021; Singh et al. 2022; Petek et al. 2022; Li and Liu 2021). With an alternative derivation and experimentation, we identify theoretical shortcomings that do not justify the empirical results let alone the assumed reliability in practice — it can be vital to understand to what degree the uncertainty estimations are trustworthy.

In the following, we resolve the supposedly unreasonable effectiveness and relate it to convergence patterns of the predicted uncertainties. Furthermore, we propose re-definitions of the original approach and discuss generalizations.

1.1 Disentangling Aleatoric and Epistemic Uncertainties

The residual between the genuine and the predicted value of an imperfect model can be disentangled into an *aleatoric* and *epistemic* contribution. In theory, the former is related to the noise level in the data and, typically, does not depend on the sample size. In contrast, the latter scales with the sample size, and either allows the model to be pulled towards the observed distribution if the sample size is increased in this region, or allows the fit of a more complex model and thus decreasing under-fitting in general. In practice, however, both types can be non-trivially correlated and disentangling both without relying on a-priori assumptions is often ambiguous. Even if both types are uncorrelated, observing a single de-

*These authors contributed equally.

viation from the ground truth $\delta(x)$ at point x, linear error propagation, $\delta^2(x) = u_{\text{al}}^2(x) + u_{\text{ep}}^2(x)$, shows that a separation of the aleatoric $u_{\text{al}}(x)$ and the epistemic uncertainty $u_{\text{ep}}(x)$ is impossible to do point-wise without assuming a certain level of smoothness in x and prior knowledge about at least one of the uncertainties.

For now take the simplest setting: univariate, point-wise normal distributed data, i.e., $(x_i, y_i + \epsilon_i)$ where the noise, $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, is heteroscedastic, i.e., not necessarily equally distributed for all x_i and our goal is to predict the mean and the noise level depending on the input x_i. In a Bayesian framework, this corresponds to taking a normal-inverse-gamma distribution, $\text{NIG}(\mu, \sigma^2|\boldsymbol{m})$ with $\boldsymbol{m} = (\gamma, \nu, \alpha, \beta)$, as the conjugated prior of a normal distribution with unknown mean μ and variance σ^2. Integrating out the nuisance parameters, Bayesian inference yields that the likelihood of an observation y_i given $\boldsymbol{m}$ follows a t-distribution with $2\alpha_i$ degrees of freedom (Gelman et al. 2013; Amini et al. 2020),

$$L_i^{\text{NIG}} = \text{St}_{2\alpha_i}\left(y_i \middle| \gamma_i, \frac{\beta_i(1+\nu_i)}{\nu_i \alpha_i}\right). \qquad (2)$$

If $\boldsymbol{m}$ is known, it is reasonable to define the prediction of y_i as $\mathbb{E}[\mu_i] = \gamma_i$, and the aleatoric and epistemic uncertainties u_{al} and u_{ep}, respectively, as:

$$u_{\text{al}}^2 \equiv \mathbb{E}[\sigma_i^2] = \beta_i/(\alpha_i - 1) \quad u_{\text{ep}}^2 \equiv \text{var}[\mu_i] = \mathbb{E}[\sigma_i^2]/\nu_i\,. \qquad (3)$$

1.2 State of the Art and Its Issues

One may learn Eq. (2) from data by minimizing

$$\mathcal{L}_i(\boldsymbol{\omega}) = -\log L_i^{\text{NIG}}(\boldsymbol{\omega}) + \lambda \underbrace{|y_i - \gamma_i|\Phi}_{\mathcal{L}_i^{\text{R}}(\boldsymbol{\omega})}, \qquad (4)$$

where $\boldsymbol{m} = \text{NN}(\boldsymbol{\omega})$ is given by a NN, λ is a tunable hyperparameter, and $\Phi = 2\nu_i + \alpha_i$ is the *total evidence*. The definition of the total evidence is motivated by the fact that taking a NIG distribution as a conjugated prior corresponds to assuming prior knowledge about the mean and the variance extracted from ν_i virtual measurements of the former and $2\alpha_i$ virtual measurements for the latter (Gelman et al. 2013). For consistency we have adopted the definition of Φ as shown above but note that using $\Phi = \nu_i + 2\alpha_i$ would actually be better motivated as pointed out first by Meinert and Lavin (2021) and already used as such in Liu et al. (2021).

The issue with the existing approach, as proposed by Amini et al. (2020) and shown in Eq. (4), is that minimizing $\mathcal{L}_i(\boldsymbol{\omega})$ w.r.t. $\boldsymbol{\omega}$ is insufficient to find $\boldsymbol{m}$. This is obvious by noting the overparameterization in Eq. (2) that is not resolved by adding the regularization $\mathcal{L}_i^{\text{R}}(\boldsymbol{\omega})$, and can also be shown with some mathematical rigor by finding

$$\frac{\partial}{\partial \nu_i} \log L_i^{\text{NIG}} = 0 \quad \text{if}\, \beta_i(\nu_i) \propto \frac{1}{1 + \nu_i^{-1}}, \qquad (5)$$

to reveal that $\log L_i^{\text{NIG}}$ does not depend on ν_i and thus $\mathcal{L}_i(\boldsymbol{\omega})$ is minimized, independent of the data, by sending $\nu_i \to 0$ and, for instance, following a path along $\beta_i(\nu_i) = 1/(1 + \nu_i^{-1})$. Clearly, a loss function which can be minimized independent of data is non-informative for $\boldsymbol{m}$ and thus cannot be used for evaluating Eqs. (3). The underlying reason for this overparameterization, which allows for choosing a path which minimizes the loss independent of data, is the fact that L_i^{NIG} is by definition a projection of the NIG distribution and thus unable to unfold all of its degrees of freedom unambiguously. This ambiguity stays unresolved due to a missing additional constraint for β_i in the regularizer $\mathcal{L}_i^{\text{R}}$.

Curiously, if applied to synthetic and real-world data, the aforementioned approach does yield reasonable results (Amini et al. 2020; Liu et al. 2021; Cai et al. 2021; Soleimany et al. 2021; Singh et al. 2022; Petek et al. 2022; Li and Liu 2021). In particular, areas with a low data density during training that result in a large model uncertainty during inference are identified with large values in the epistemic uncertainty if estimated according to Eq. (3) and low values elsewhere.

2 The Unreasonable Effectiveness

In this section we show why Deep Evidential Regression (DER) can yield reasonable results in practice despite its issues — i.e., why it is unreasonably effective — and we identify and redefine the key components of extracting disentangled uncertainties.

2.1 Measuring Aleatoric Uncertainty by Convergence Speed

In order to analyze the convergence behavior of DER we set up the same synthetic experiment that was used by Amini et al. (2020). A shallow, fully connected NN with a single input and four output neurons, $(\theta_{i,1}, \theta_{i,2}, \theta_{i,3}, \theta_{i,4})$, is used to predict for a given x_i the parameters of a NIG distribution $\boldsymbol{m} = (\gamma_i, \nu_i, \alpha_i, \beta_i)$,

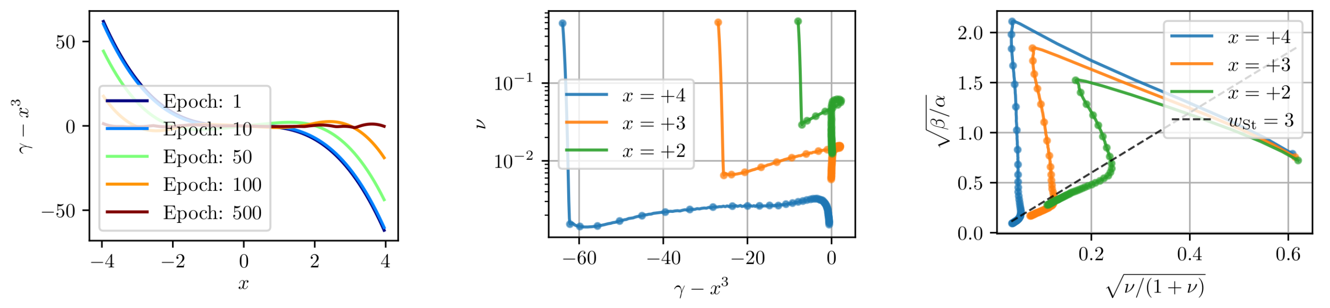
$$\begin{aligned} \gamma_i &= \theta_{i,1}, & \nu_i &= \text{softplus}(\theta_{i,2}), \\ \alpha_i &= \text{softplus}(\theta_{i,3}) + 1, & \beta_i &= \text{softplus}(\theta_{i,4}), \end{aligned} \qquad (6)$$

where $\text{softplus}(\bullet) = \log(1 + \exp(\bullet))$ enforces non-negative values.

A data sample is generated by generating 1k pairs $(x_i, x_i^3 + \epsilon_i)$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2 = 9)$ and x_i uniformly distributed on $[-4, +4]$. The NN is trained with $\lambda = 0.01$ and the Adam optimizer with an initial learning rate of 5×10^{-4} for 500 epochs. We repeat this experiment 50 times with different seeds for the initialization of the NN. The results of multiple independently trained NNs are shown in the Appendix (see Sec. 6) and reproduce the findings of Amini et al. (2020) despite a certain variance among the samples: On the larger interval $x \in [-7, +7]$, the model predicts a large epistemic uncertainty in regions where it has never seen data during training.

In Fig. 1a we show the convergence of $\gamma_i(x)$ in time and find that the convergence speed differs across the interval in x. Large residuals induce large gradients for ν_i in the regularizer,

$$\frac{\partial \mathcal{L}_i^{\text{R}}}{\partial \nu_i} \propto \lambda |y_i - \gamma_i|, \qquad (7)$$



(a) Evolution of residual.

(b) Evolution of ν w.r.t. residual.

(c) Evolution of factors of w_{St}.

Figure 1: Evolution of parameters during training. In total, 50 independently trained samples are averaged in each epoch. Dots are placed every 10 epochs to indicate the speed and direction of the convergence. (Left) Evolution of the residual $\gamma_i - x^3$ for all $x \in [-4, 4]$. (Center) Evolution of ν_i w.r.t. residual at three different x-positions. (Right) Evolution of $\sqrt{\beta_i/\alpha_i}$ and $\sqrt{\nu_i/(1+\nu_i}$ at different x positions. The ratio of the quantities is the estimation of the width of a t-distribution, w_{St}. A constant width $w_{\mathrm{St}} = 3$ is indicate by a dashed line and referred to in the text as the *valley*.

and thus push ν_i towards smaller values faster, as shown in Fig. 1b. Our studies show that residuals of the very first epochs predominantly define how small ν_i will eventually become since the magnitude of its gradient gradual decreases the better the model gets.

In the Appendix we show the corresponding analysis for α_i and find a similar behavior, however, here a second gradient from $\log L_i^{\mathrm{NIG}}$ is conflicting with the gradient of the regularizer which slows down convergence and eventually stop at values $\alpha_i \approx 2$ for $-4 < x_i < +4$.

In summary: our first key insight to understand DER is that the point-wise convergence speed of the model is used as a proxy for the epistemic uncertainty, which arguably stretches the canonical definition. As a consequence, the numerical values of the epistemic uncertainty cannot be interpreted as canonical Bayesian or Frequentist uncertainty estimations and only relative changes are conclusive. In that sense, DER is a heuristic that has proven to yield decent results in practice.

2.2 How To Not Extend DER

It is tempting to generalize this approach and use it for arbitrary functions $f(\vec{x}|\boldsymbol{\omega})$, that sufficiently describes the data distribution, by adding a regularizer to the NLL part of the loss function,

$$\mathcal{L}_i(\boldsymbol{\omega}, \nu_i) = -\log f(\boldsymbol{\omega}) + \lambda|\vec{y}_i - \vec{\gamma}_i|\nu_i. \tag{8}$$

However, empirically we find that optimizers, especially those that include momentum, will rapidly push $\nu_i \approx 0$ within machine precision even for very small coupling constants λ, rendering the proxy $\sim 1/\sqrt{\nu_i}$ useless. Contrarily, in Fig. 1b we see that a fast drop is stopped in DER after a few epochs. The reason for this can be seen in Fig. 1c where we show the evolution of $\sqrt{\beta_i/\alpha_i}$ and $\sqrt{\nu_i/(1+\nu_i)}$ which together are the width w_{St} of the t-distribution,

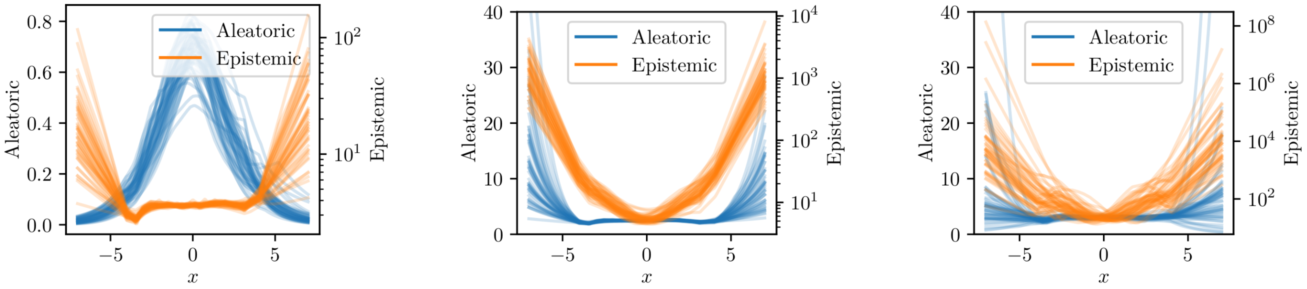
$$w_{\mathrm{St}} = \sqrt{\frac{\beta_i(1+\nu_i)}{\alpha_i \nu_i}}. \tag{9}$$

It is straightforward to show that for large deviations, the gradients for γ_i and w_{St} push both towards larger values and are therefore aligned with decreasing values for ν_i. This is the second key insight to understand DER: Optimizers are confused by the non-trivial correlation[1] in w_{St} with the parameters β_i and α_i which effectively defers the minimization goal induced by the regularizer. Once the gradient of w_{St} has flipped sign, it depends on the coupling constant λ how stiff ν_i is kept until the correct width, $w_{\mathrm{St}} \approx 3$, is reached. Only then, the optimizer starts to follow the *valley* in the loss function (dashed line in Fig. 1c) that simultaneously preserves w_{St} but decreases the total evidence. Here, the residual is already small which limits the gradient induced by the regularizer.

2.3 Redefining Proxies for Aleatoric and Epistemic Uncertainties

We have shown above that the derivation of Eq. (4) does not help to obtain the correct aleatoric or epistemic uncertainties in a strict Bayesian sense. The fact that also the aleatoric prediction is spoiled follows directly by the insights we have described previously, but can also be seen visually by plotting $u_{\mathrm{al}} = \sqrt{\mathbb{E}[\sigma_i^2]}$ as a function of x: In Fig. 2a the prediction of the aleatoric uncertainty materializes as a peaking structure with a peak height of roughly 0.7, whereas the data were generated with a constant standard deviation of $\sigma = 3$ for all x. Unsurprisingly and in accordance with Fig. 1c, this value is fitted in good approximation by the width of the t-distribution, w_{St}, due to the close resemblance between a t-distribution and a normal distribution. Intuitively, the standard deviation of a normal distribution, approximated by the w_{St}, can be interpreted as the aleatoric uncertainty of the data

[1] Jacobian based optimizers cannot exactly follow curved gradients as induced by non-trivial correlations but approximately follow with piece-wise linear segments. As a consequence, optimizers overshoot and do not accumulate large momentum in these scenarios.



(a) SOTA

(b) Proposed

(c) Alternative

Figure 2: We propose a redefinition of the aleatoric and epistemic uncertainty. (Left) The prediction for both uncertainties is shown using the SOTA definitions after training the NN for 500 epochs. The training is repeated with 50 different seeds. (Center) Results of the same NN but the uncertainties are estimated by our proposed definition in Eq. (10). Note that most of the characteristic features of the epistemic uncertainty is shifted into the aleatoric uncertainty by our redefinition. (Right) Results of using the modified loss in Eq. (12).

and we therefore propose to redefine Eq. (3) accordingly. Doing so unveils that the plateau between $-4 < x < +4$ of $\mathbb{E}[\sigma_i^2]$ was mainly a characteristic feature of the aleatoric and not the epistemic uncertainty! In fact, it is reasonable to adopt the previous relation between aleatoric and epistemic uncertainty and redefine the latter as $1/\sqrt{\nu_i}$ as shown in Fig. 2b, which makes our proposed proxies for the aleatoric uncertainty u'_{al} and the epistemic uncertainty u'_{ep}:

$$u'_{\text{al}} \equiv w_{\text{St}} = \sqrt{\frac{\beta_i(1+\nu_i)}{\alpha_i \nu_i}} \qquad u'_{\text{ep}} \equiv \frac{u_{\text{ep}}}{u_{\text{al}}} = \frac{1}{\sqrt{\nu_i}}\,. \tag{10}$$

Finally, we note that scaling the gradient of ν_i by the magnitude of the residual, regions with a high noise level also tend to contribute large gradients. A possibility to disentangle this aleatoric contribution is to normalize the residual with w_{St},

$$\mathcal{L}_i(\boldsymbol{\omega}) = -\log L_i^{\text{NIG}}(\boldsymbol{\omega}) + \lambda \left|\frac{y_i - \gamma_i}{w_{\text{St}}}\right|^p \Phi. \tag{11}$$

Most importantly, this inhibits the convergence for large, yet insignificant residuals in terms of the associated aleatoric uncertainty, in particular when the fit has reached the *valley*.

For the synthetic example $(x_i, x_i^3 + \epsilon_i)$, we test the proposed change with $p = \{1, 2\}$ and do not find any significant deviations w.r.t. the original formulation of the regularizer. This is in good agreement with our expectation since in this example the variance of ϵ_i is constant over the generated sample. However, in Sec. 3.1 we test our proposed loss function on a data set with varying noise level and find that Eq. (11) more effectively strips this aleatoric component from the predicted epistemic uncertainty than DER with Eq. (4).

2.4 Generalization and Its Limitations

In the Bayesian framework the parameter α_i is associated with the number of virtual measurements encoded in the prior. In the limit $\alpha_i \to \infty$, the t-distribution with $2\alpha_i$ degrees of freedom becomes the normal distribution that was used for generating the synthetic data. Naïvely, one would therefore expect this number to be large at places with high model accuracy and data density, but, as we have eluded before, we find values $\alpha_i \approx 2$ instead. In practice, this discrepancy causes a small offset between w_{St} and the genuine standard deviation of the noise level. Also, we find it noteworthy to mention that even if the data were genuinely distributed according to the t-distribution, the strong correlation of α_i and w_{St} makes a clean extraction with a fitting approach challenging.

This observation motivates an extension of the mean-variance estimation by Nix and Weigend (1994) as a generalization of DER:[2] The likelihood of the t-distribution is replaced with a normal distribution $\mathcal{N}(y_i|\gamma_i, \sigma_i^2 = \beta_i/\nu_i)$ where β_i is used to slow down the convergence of ν_i.

Using our ansatz in Eq. (11) with $p = 2$, the alternative loss functions reads

$$\mathcal{L}'_i = \log \sigma_i^2 + (1 + \lambda\nu_i)\,\frac{(y_i - \gamma_i)^2}{\sigma_i^2}. \tag{12}$$

Similarly, other data distributions, such as a Poisson or a Log-normal distribution, could be adopted by replacing the PDF of the normal distribution accordingly. Clearly, this generalization only affects the aleatoric uncertainty estimation whereas the proxy for the epistemic uncertainty remains the same. It is therefore pivotal to analyze, if the separation of the uncertainty types by stripping α_i from the equation preserves the good properties we have witnessed before.

We take our previous example[3] and use the modified loss function here. We use the same NN, ignore $\theta_{i,3}$, and train it 50 times with $\lambda = 2$ and a learning rate of 5×10^{-3} for 500 epochs. The uncertainty predictions shown in Fig. 2c look similar, yet the sample variance is significantly higher.

[2] Unlike Nix and Weigend (1994), we learn all parameters simultaneously.

[3] We also have used Eq. (12) for our example in Sec. 3.1.

This is what we finally coin the *unreasonable effectiveness* of DER. Defining the total evidence by two, rather than a single parameter, keeps the optimizer sufficiently busy during optimization and prevents a too fast convergence, thus reducing the dependency to the random initialization of ν_i and α_i. Furthermore, coupling α_i non-trivially in the NLL part of the loss function stabilizes the convergence even more. This comes by the price of a bias due to the approximation of the standard deviation of the noise level with the width of a t-distribution. For most applications the smaller sampling variance should offset this disadvantage, though.

3 Experiments

3.1 Binary Pulse

The original formulation of DER uses the residual $|y_i - \gamma_i|$ to scale the gradients of the total evidence Φ. Apparently, this approach works fine if the variance of the noise level is constant for all training data. If this is not the case, though, parts of the aleatoric uncertainty leak asymmetrically into the estimation of the epistemic uncertainty. We demonstrate this with a simple synthetic data sample $(x_i, y_i + \epsilon_i)$ with a point-wise normal distributed noise $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and

$$y_i = \begin{cases} 1 & \text{if } |x_i - .5| < .0025 \\ 0 & \text{else} \end{cases} \qquad \sigma_i^2 = \begin{cases} 10^{-4} & \text{if } x_i < .5 \\ 10^{-2} & \text{else.} \end{cases} \tag{13}$$

We sample 1k data points and train the same shallow NN used previously, trained with $\lambda = 0.01$ and a learning rate of 1×10^{-3} for 600 epochs. The data distribution and the corresponding predictions of the NN are shown in Fig. 3a.

The predicted uncertainties according to Eqs. (10) when trained with Eq. (4) and Eq. (12) are shown in Fig. 3b and Fig. 3c, respectively.

The synthetic data sample was chosen such that the model significantly under-fits the peak region and we expect to see a large epistemic uncertainty here. Obviously, also the aleatoric uncertainty is pulled toward the peak in this region which is reasonable since large values of σ_i^2 indeed improve the total loss if the deviation is large. More importantly, we see that with our modified version not only the aleatoric uncertainty prediction is better but also the epistemic uncertainty is symmetric, despite a certain sampling variance we have already witnessed before.

This example points towards a useful interpretation of u'_{al} and u'_{ep} from Eqs. (10), in particular if paired with the modified loss (12): The former is the point-wise prediction of the standard deviation of an additive, normal distributed noise source in the data, and the latter indicates when the estimation of the former is vague.

3.2 Monocular Depth Estimation

Finally, we reevaluate the performance of DER on the same, high-dimensional task of depth estimation that was used in the prior art. The training set consists of over 27k RGB-to-depth image pairs of indoor scenes from the NYU Depth v2 dataset (Silberman et al. 2012). We use the scripts[4] for training and testing provided by Amini et al. (2020), and only replace the actual estimation of u_{ep} by applying the patches printed in the Appendix.

Following Kendall and Gal (2017); Kuleshov, Fenner, and Ermon (2018); Amini et al. (2020), we show in Fig. 4a how DER performs as pixels with uncertainty greater than certain thresholds are removed when the uncertainty is predicted by u_{ep}, u'_{ep} or u'_{al}. In full agreement with our previous findings, e.g., as shown in Fig. 2, our redefinition of the epistemic uncertainty, u_{ep}, shifts a significant part into the aleatoric uncertainty, u'_{al}. The same behavior is seen in the calibration curves[5] shown in Fig. 4b. Our redefined proxy for the epistemic uncertainty, u'_{ep}, appears less helpful whereas $u_{\text{ep}} \approx u'_{\text{al}}$ in good approximation.

Since this is a complex data set we can only speculate about the reason for the apparent effectiveness of estimating epistemic uncertainty with a proxy for aleatoric uncertainty: Still, in our previous experiments we already saw and described the phenomenon that the width of the t-distribution (or, similarly, the standard deviation of a normal distribution) tends to massively overshoot if the genuine model uncertainty is large. In our experiments this overshooting was much larger than the genuine noise level of the data, hence, macroscopically and only if the genuine aleatoric uncertainty is sufficiently small, u'_{al} indeed is a proxy for the epistemic uncertainty.

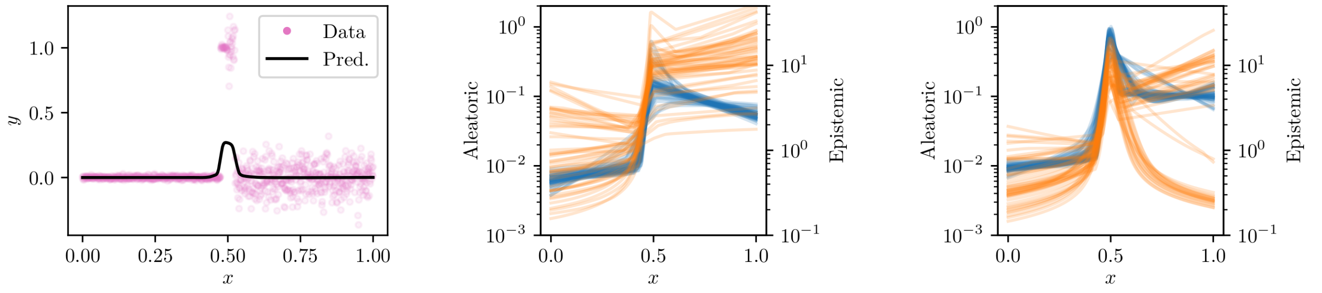
To verify our assumption we replace the loss function with Eq. (12) and set $\lambda = 0$, and obtain the canonical NLL of a normal distribution as discussed by Nix and Weigend (1994). The results of using the standard deviation of the normal distribution after 26k and 45k iterations as the proxy for the epistemic uncertainty are shown in Figs. 4 and Fig. 4c. We observe a shift of a well-calibrated aleatoric proxy towards a less well calibrated, yet more powerful separator of in- and out-of-distribution data and anticipate that a similar transformation also leverages the good performance of DER in this task.

4 Conclusion and Outlook

Deep Evidential Regression (DER) is frequently used in the field as a tool to disentangle aleatoric and epistemic uncertainties. We theoretically showed that the representation of the uncertainties is over parametrized making the efficiency of the DER seam almost unreasonable. Using synthetic data, we demystified this unreasonable effectiveness of DER and related it to the convergence patterns of the learned uncertainty representations during the training process. We also demonstrated that the estimation of the aleatoric uncertainty as used in the prior art yields quantitatively and qualitatively wrong results which brings us to the conclusion that w_{St} does not disentangle aleatoric and epistemic uncertainties, however, it represents a reasonable proxy of the later in practice.

[4] https://github.com/aamini/evidential-deep-learning, git commit hash: d1d8e39

[5] For each predicted μ_i and σ_i, the inverse CDF of a normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ is used to get the upper boundary of a confidence interval. The fraction of all predictions below this value is the observed CL. See the work of Kendall and Gal (2017); Kuleshov, Fenner, and Ermon (2018) for more details.

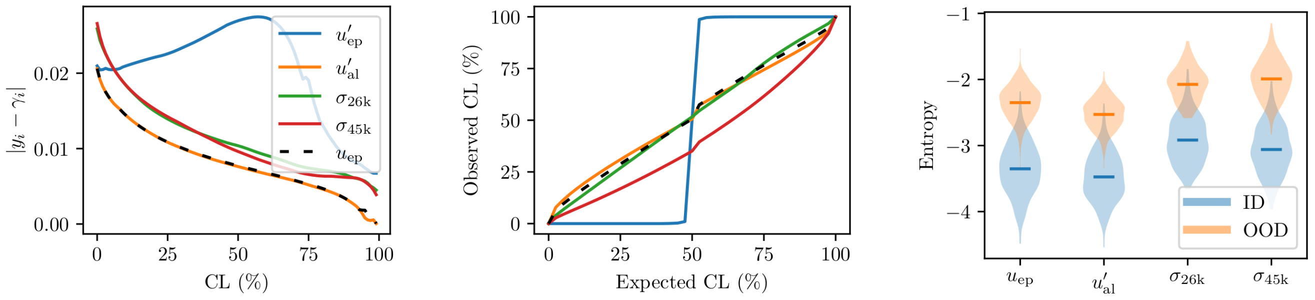


(a) Data and prediction.

(b) SOTA

(c) Proposed

Figure 3: Results of the binary pulse experiment. (Left) The averaged prediction γ_i for our synthetic data sample as described in Eqs. (13). (Center) Estimations of the aleatoric (blue) and epistemic (orange) uncertainty using a regularizer proportional to the residual as used in the prior art. (Right) Uncertainty estimation using a normalized residual as proposed in Eq. (12).



(a) Prediction CL vs. observed error.

(b) Uncertainty calibration.

(c) Violin plots of the entropy.

Figure 4: (Left) Relationship between prediction CL and observed error. (Center) Model uncertainty calibration if u_{ep} is replaced with u'_{ep} or u'_{al}. A strong inverse trend for the former is desired. An ideal calibration is when the expected and the observed CL match. In addition, we also show the results of training a NLL of a normal distribution and using its standard deviation, σ, after 26k and 45k iterations. (Right) Violin plots of the entropy, $\log(2\pi\sigma^2)/2$, where $\sigma = \{u_{\text{ep}}, u'_{\text{al}}, \sigma_{26\text{k}}, \sigma_{45\text{k}}\}$ is the respective prediction of the epistemic uncertainty. We compare the results of in-distribution data (ID) and out-of-distribution data (OOD) where the latter were not part of the training. The OOD data are taken from Huang et al. (2018).

In our experiments with monocular depth estimation, we showed that the estimation of the epistemic uncertainty in the prior art (Amini et al. 2020) is nearly equivalent to fitting the width of the t-distribution, w_{St}, a quantity that is, naïvely, tightly coupled with the aleatoric uncertainty, missing the main disentanglement goal.

As a main result of this work we show that the behavior of DER can be understood with a fit of a Gaussian NLL, albeit, DER defers convergence speed such that w_{St} becomes almost unreasonably effective. Practitioners should treat DER more as a heuristic than an exact uncertainty quantification and carefully validate its predictions, while using the redefinitions and investigative studies we described here for disentangling uncertainties in NNs.

However, our results are negative in the following sense: Although helpful in understanding the genuine meaning of DER predictions, we cannot fundamentally fix the underlying issue and deliver what was promised in the work of Amini et al. (2020). Instead, we like to motivate to investigate, given a specific application, whether it is worthwhile to strive for a precise epistemic uncertainty in the first place: As pointed out by Bengs, Hüllermeier, and Waegeman (2022), epistemic uncertainty is difficult to quantify objectively, because, unlike aleatoric uncertainty, epistemic uncertainty is not a property of the data or the data-generating process, and there is nothing like a *ground truth* epistemic uncertainty. If the consequence of a large epistemic uncertainty is to reject a model decision entirely and to fall back to, e.g., human supervision, the nominal value of the predicted uncertainty becomes less important and heuristics, such as DER, that do not require OOD data during training to supervise instances of high uncertainty, are actually still highly valuable.

More drastically, our findings of Sec. 1.2 that a projection cannot be used to unfold all degrees of freedom unambiguously might hint toward a similar fundamental problem that was reported recently by Bengs, Hüllermeier, and Waegeman (2022); Hüllermeier (2022). As a consequence, trying to exactly disentangle different types of uncertainties with *second-order learners* with purely loss-based methods (e.g., evidential NNs) might be impossible and the Bayesian derivation of Eq. (2) is one way to prove it.

Finally, our observation of convergence speed as a proxy might circumvent the general problem with second-order learners and purely loss-based approaches that were studied by Bengs, Hüllermeier, and Waegeman (2022). Indeed, the convergence history can extract valuable information and, in a sense, does scale with the amount of information shown as labeled data during each epoch of a training sequence. Hypothetically, our observation could be extended in future works to improve uncertainty prediction for regression but also for classification tasks and other related fields.

5 Related Work

Our work builds on the prior art (Amini et al. 2020) for uncertainty estimation with evidential neural networks and the technical report of Meinert and Lavin (2021), and more generally on the advancing area of uncertainty reasoning in deep learning.

In recent years there have been many explorations into Bayesian approaches to deep learning (Kendall and Gal 2017; Neal 1996; Guo et al. 2017; Wilson et al. 2016; Hafner et al. 2019; Ovadia et al. 2010; Izmailov et al. 2019; Seedat and Kanan 2020). The key observation is that neural networks are typically underspecified by the data, thus different settings of the parameters correspond to a diverse variety of compelling explanations for the data — i.e., a deep learning posterior consists of high performing models which make meaningfully different predictions on test data, as demonstrated by Izmailov et al. (2019); Garipov et al. (2018); Zolna, Geras, and Cho (2020). This underspecification by NNs makes Bayesian inference, and by corollary uncertainty estimation, particularly compelling for deep learning. Bayesian deep learning aims to compute a distribution over the model parameters during training in order to quantify uncertainties, such that the posterior is available for uncertainty estimation and model calibration (Guo et al. 2017). With Bayesian NNs that have thousands and millions of parameters this posterior is intractable, so implementations largely focus on several approximate methods for Bayesian inference: First, Markov Chain Monte Carlo (MCMC) methods and in particular stochastic gradient MCMC for Bayesian NNs (Welling and Teh 2011; Li et al. 2016; Park et al. 2018; Maddox et al. 2019) show promise, with a main drawback being the inability to capture complex distributions in the parameter space without increasing the computational overhead. Secondly, variational inference (VI) performs Bayesian inference by using a computationally tractable *variational* distribution to approximate the posterior. One approach by Graves, Mohamed, and Hinton (2013) is to use a Gaussian variational posterior to approximate the distribution of the weights in a network, but the capacity of the uncertainty representation is limited by the variational distribution. In general we see that MCMC has a higher variance and lower bias in the estimate, while VI has a higher bias but lower variance (Mattei 2020). The preeminent Bayesian deep learning approach by Gal and Ghahramani (2016) showed that variational inference can be approximated without modifying the network. This is achieved through a method of approximate variational inference called Monte Carlo Dropout (MCD), whereby dropout is performed during inference, using multiple forward passes with randomly sampled dropout masks. Kendall and Gal (2017) used a combination of mean-variance estimation (Nix and Weigend 1994) and MCD to simultaneously predict aleatoric and epistemic uncertainties. However, this approach is limited by its requirement of multiple forward passes to gather enough information for a decent approximation by MCD which often makes this method uneconomical in practical applications.

Alternative to the prior-over-weights approach of Bayesian NN, one can view deep learning as an evidence acquisition process — different from the Bayesian modeling nomenclature, evidence here is a measure of the amount of support collected from data in favor of a sample to be classified into a certain class, and uncertainty is inversely proportional to the total evidence (Sensoy, Kaplan, and Kandemir 2018). Samples during training each add support to a learned higher-order, evidential distribution, which yields epistemic and aleatoric uncertainties without the need for sampling. Several recent works develop this approach to deep learning and uncertainty estimation which put this in practice with *prior networks* that place priors directly over the likelihood function (Amini et al. 2020; Malinin and Gales 2018). These approaches largely struggle with regularization (Sensoy, Kaplan, and Kandemir 2018), generalization (particularly without using out-of-distribution training data) (Malinin and Gales 2018; Hafner et al. 2019), capturing aleatoric uncertainty (Gurevich and Stuke 2019), and the issues we have addressed above with the prior art Deep Evidential Regression (Amini et al. 2020).

There are also the frequentist approaches of bootstrapping and ensembling, which can be used to estimate NN uncertainty without the Bayesian computational overhead as well as being easily parallelizable — for instance Deep Ensembles, where multiple randomly initialized NNs are trained and at test time the output variance from the ensemble of models is used as an estimate of uncertainty (Lakshminarayanan, Pritzel, and Blundell 2017).

6 Reproducibility

An open-source GitHub repository (MIT License) with the Appendix and source code for reproducing our experiments (implementations of the NNs and algorithms to generate the data) is available on https://github.com/pasteurlabs/unreasonable_effective_der. We encourage other researchers to reproduce, test, extend, and apply our work. The model and experiments are lightweight, running locally on a 4-core MacBook Pro in under an hour.

References

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D. 2020. Deep Evidential Regression. *Advances in Neural Information Processing Systems*, 33.

Bengs, V.; Hüllermeier, E.; and Waegeman, W. 2022. Pitfalls of Epistemic Uncertainty Quantification through Loss

Minimisation. *Advances in Neural Information Processing Systems*, 35.

Cai, P.; Wang, H.; Huang, H.; Liu, Y.; and Liu, M. 2021. Vision-Based Autonomous Car Racing Using Deep Imitative Reinforcement Learning. *IEEE Robotics and Automation Letters*, 6(4).

Chen, X.; Bromuri, S.; and van Eekelen, M. 2021. *Neural Machine Translation for Harmonized System Codes Prediction*. Association for Computing Machinery. ISBN 978-1-450-38940-2.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *International Conference on Machine Learning*, 48.

Garipov, T.; Izmailov, P.; Podoprikhin, D.; Vetrov, D.; and Wilson, A. G. 2018. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. *Advances in Neural Information Processing Systems*, 31.

Gelman, A.; Carlin, J.; Stern, H.; Dunson, D.; Vehtari, A.; and Rubin, D. 2013. *Bayesian Data Analysis*. Taylor & Francis Ltd. ISBN 978-1-439-84095-5.

Graves, A.; Mohamed, A.; and Hinton, G. 2013. Speech Recognition with Deep Recurrent Neural Networks. *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. *International Conference on Machine Learning*, 34.

Gurevich, P.; and Stuke, H. 2019. Gradient conjugate priors and multi-layer Neural Networks. *Artificial Intelligence*, 278.

Hafner, D.; Tran, D.; Lillicrap, T.; Irpan, A.; and Davidson, J. 2019. Noise Contrastive Priors for Functional Uncertainty. *Proceedings of Machine Learning Research*, 115.

Hora, S. C. 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2-3).

Huang, X.; Cheng, X.; Geng, Q.; Cao, B.; Zhou, D.; Wang, P.; Lin, Y.; and Yang, R. 2018. The ApolloScape Dataset for Autonomous Driving. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Hüllermeier, E. 2022. Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning: Are Conditional Entropy and Mutual Information Appropriate Measures? *arXiv:2209.03302 [cs.LG]*.

Hüllermeier, E.; and Waegeman, W. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110.

Izmailov, P.; Maddox, W. J.; Kirichenko, P.; Garipov, T.; Vetrov, D. P.; and Wilson, A. G. 2019. Subspace Inference for Bayesian Deep Learning. *Proceedings of Machine Learning Research*, 115.

Kendall, A.; and Gal, Y. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in Neural Information Processing Systems*, 30.

Kiureghian, A. D.; and Ditlevsen, O. 2009. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2).

Kuleshov, V.; Fenner, N.; and Ermon, S. 2018. Accurate Uncertainties for Deep Learning Using Calibrated Regression. *International Conference on Machine Learning*, 35.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Advances in Neural Information Processing Systems*, 30.

Li, C.; Stevens, A.; Chen, C.; Pu, Y.; Gan, Z.; and Carin, L. 2016. Learning Weight Uncertainty with Stochastic Gradient MCMC for Shape Classification. *IEEE Conference on Computer Vision and Pattern Recognition*.

Li, H.; and Liu, J. 2021. 3D High-Quality Magnetic Resonance Image Restoration in Clinics Using Deep Learning. *arXiv:2111.14259 [eess.IV]*.

Liu, Z.; Amini, A.; Zhu, S.; Karaman, S.; Han, S.; and Rus, D. 2021. Efficient and Robust LiDAR-Based End-to-End Navigation. *IEEE International Conference on Robotics and Automation*.

Maddox, W.; Garipov, T.; Izmailov, P.; Vetrov, D.; and Wilson, A. G. 2019. A Simple Baseline for Bayesian Uncertainty in Deep Learning. *Advances in Neural Information Processing Systems*, 32.

Malinin, A.; and Gales, M. 2018. Predictive Uncertainty Estimation via Prior Networks. *Advances in Neural Information Processing Systems*, 31.

Mattei, P.-A. 2020. A Parsimonious Tour of Bayesian Model Uncertainty. *arXiv:1902.05539 [stat.ME]*.

Meinert, N.; and Lavin, A. 2021. Multivariate Deep Evidential Regression. *arXiv:2104.06135 [cs.LG]*.

Neal, R. M. 1996. *Bayesian Learning for Neural Networks*. Springer. ISBN 978-1-4612-0745-0.

Nix, D. A.; and Weigend, A. S. 1994. Estimating the mean and variance of the target probability distribution. *IEEE International Conference on Neural Networks*.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J. V.; Lakshminarayanan, B.; and Snoek, J. 2010. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *Advances in Neural Information Processing Systems*, 32.

Park, C.; Kim, J. M.; Ha, S. H.; and Lee, J. 2018. Sampling-based Bayesian Inference with gradient uncertainty. *NeurIPS Workshop on Bayesian Deep Learning*.

Petek, K.; Sirohi, K.; Büscher, D.; and Burgard, W. 2022. Robust Monocular Localization in Sparse HD Maps Leveraging Multi-Task Uncertainty Estimation. *International Conference on Robotics and Automation*, 39.

Seedat, N.; and Kanan, C. 2020. Towards calibrated and scalable uncertainty representations for Neural Networks. *Advances in Neural Information Processing Systems*, 33.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. *Advances in Neural Information Processing Systems*, 31.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor Segmentation and Support Inference from RGBD Images. *European Conference on Computer Vision*.

Singh, S. K.; Fowdur, J. S.; Gawlikowski, J.; and Medina, D. 2022. Leveraging Evidential Deep Learning Uncertainties with Graph-based Clustering to Detect Anomalies. *IEEE Transactions on Intelligent Transportation Systems*, 25.

Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; and Coley, C. W. 2021. Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Central Science*, 7(8).

Welling, M.; and Teh, Y. W. 2011. Bayesian learning via stochastic gradient langevin dynamics. *International Conference on Machine Learning*, 28.

Wilson, A. G.; Hu, Z.; Salakhutdinov, R.; and Xing, E. P. 2016. Deep Kernel Learning. *International Conference on Artificial Intelligence and Statistics*, 19.

Zolna, K.; Geras, K. J.; and Cho, K. 2020. Classifier-agnostic saliency map extraction. *Computer Vision and Image Understanding*, 196.